IEEE TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editör Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

IEEE TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

Media type	Part Number	ISBN	Online ISSN
XPLORE COMPLIANT	CFP25L97-ART	979-8-3315-9975-1	2521-1641
CD-ROM	CFP25L97-CDR	979-8-3315-9974-4	

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK’2025)

10th International Conference on Computer Science and Engineering

17-18-19 Eylül 2025 -İstanbul-Türkiye
17-18-19 September 2025 - İstanbul-Türkiye

IEEE Catalog Number : CFP25L97-ART

ISBN : 978-8-3315-9975-1

Additional copies mey be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

UBMK’2025’ye Hoşgeldiniz
Welcome to UBMK’2025

Sevgili Katılımcılar:

UBMK uluslararası nitelikli konferans serisi, 1990 yılından beri düzenli olarak yapılmakta olan Bilgisayar Mühendisliği Bölüm Başkanları toplantılarında alınan bir kararla on yıl önce başlamıştır. Konferansın 10.su IEEE-UBMK-2025 bu yıl 17-18-19 Eylül, 2025 günlerinde İstanbul Teknik Üniversitesinin ev sahipliğinde düzenlemiştir.

IEEE-UBMK-2025 konferansına bu yıl Almanya, Amerika Birleşik Devletleri, Azerbaycan, Fransa, Irak, İngiltere, İsveç, İtalya, Kanada, Kazakistan, Kırım, Kırgızistan, Rusya, Özbekistan, Tataristan, Taylant, Ürdün ve Türkiye’den 610 dolayında bildiri gönderilmiş ve bu bildiriler Türk ve yabancı 250 hakem tarafından değerlendirilmiştir.

Her bildiri en az iki hakem tarafından incelenmiş ve uzlaşma olmadığı durumlarda üçüncü bir hakemin değerlendirmesine başvurulmuştur. Bildiri başına düşen ortalama hakemlik 2,3 olmuştur. Bu değerlendirmelerin sonunda 327 bildirinin sözlü olarak sunulması uygun bulunmuştur. Kabul edilen ve sunulan bildiriler içerik ve kalite ölçünlerini sağlaması durumunda IEEE Xplore’da yayımlanacaktır.

Konferans çalışmalarında, Bilgisayar Mühendisliği Bölüm Başkanları Danışma Kurulu olarak görev almışlardır. Bildirilerin değerlendirilmesi Bilim Kurulu üyeleri tarafından yapılmıştır. Konferansın düzenlenmesi ise Yürütme Kurulunun önerileri doğrultusunda, Düzenleme Kurulu tarafından yapılmıştır.

Son olarak, konferansın başarılı bir şekilde yürütülmesi için tüm olanaklarını sunan İstanbul Teknik Üniversitesi Rektörü Sayın Prof. Dr. Hasan Mandal’a teşekkür ediyoruz. Ayrıca Düzenleme Kuruluna, bildirileri titizlikle değerlendiren Bilim Kurulu Üyelerine ve değerli araştırmalarının sonuçlarını bilişim camiası ile paylaşan bildiri sahiplerine teşekkürlerimizi iletiriz.

Prof. Dr. Eşref ADALI
UBMK-2025 Konferans Başkanı ve Bildiri Kitabı Editörü

Dear Participants:

The UBMK international conference series started nine years ago with a decision taken at the Computer Engineering Department Heads (BMBB) meetings, which have been held regularly since 1990. The 10th edition of the conference, UBMK'25, was held this year on October 17-18-19, 2025, hosted by İstanbul Technichal University.

This year, approximately 610 papers were submitted to the IEEE-UBMK-2025 conference from Germany, the United States, Azerbaijan, France, Iraq, the United Kingdom, Sweden, Italy, Canada, Kazakhstan, Crimea, Kyrgyzstan, Russia, Uzbekistan, Tatarstan, Thailand, Jordan, and Turkey, and these papers were evaluated by 250 Turkish and foreign referees.

Each paper was evaluated at least by two referees, and in cases where there was no consensus, a third referee was consulted. At the end of these evaluations, 327 papers were accepted for oral presentation. Accepted and presented papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

During the conference, Heads of Information Engineering Departments took part in the Advisory Board. The evaluation of the papers was made by the members of the Scientific Committee. The conference was organized by the Organizing Committee in line with the recommendations of the Executive Committee.

Finally, we would like to thank İstanbul Technical University Rector Prof. Dr. Hasan Mandal for his continued support for the success of the conference. In addition, we would like to thank the Organizing Committee, the Scientific Committee Members who carefully evaluated the papers, and the owners of the papers who shared the results of their valuable research with the informatics community.

Prof. Dr. Esref ADALI
UBMK'25 Conference Chair and Proceedings Editor

Düzenleyenler / Organizer

Destekleyenler / Sponsors

A Morphological Tagging Model of the Uzbek Language in the Universal Dependencies Format

Elov Botir Boltayevich
Dept.of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Khamroeva Shahlo Mirdjonovna
Dept.of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
shaxlo.xamrayeva@navoiy-uni.uz

Kurbanova Mukhabbat Matyakubovna
Department of Uzbek Linguistics, National university of Uzbekistan
Tashkent, Uzbekistan
qurbonova2007@mail.ru

Xusainova Zilola Yuldashevna
Computer Linguistics and Digital Technology
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
xusainovazilola@navoiy-uni.uz

Alavutdinova Nadira Ganiyevna
Department of Uzbek philology
National University of Uzbekistan named Mirzo Ulugbek
Tashkent, Uzbekistan
alavutdinovanodira@gmail.com

Abstract—This study presents the development of a morphological tagging model and algorithm for the Uzbek language, structured according to the Universal Dependencies (UD) framework. Taking into account the agglutinative nature of Uzbek, morphotactic rules have been systematically defined for major word classes—including nouns, verbs, adjectives, numerals, adverbs, and pronouns—using deterministic finite automata (DFA) and context-free grammars. The model specifies the class-specific sequences of affixes, their formal mapping to the UD FEATS schema, and incorporates a mathematical representation alongside a segmentation algorithm. The resulting rule-based morphological analyzer supports lemmatization, affix segmentation, and automatic grammatical tagging in compliance with UD standards. The system has been validated with over 20 example words, for which the corresponding UPOS, FEATS, and lemma annotations have been generated. Furthermore, detailed analyses of UD FEATS mapping conventions—particularly those relating to tense, mood, voice, degree, and case—have been conducted for each word class. The proposed model offers a foundational morphological component for various applications, including the development of UD-compatible corpora for Uzbek, machine translation systems, language instruction tools, and syntactic parsers.

Keywords—Universal Dependencies; morphological analysis; morphotactic model; FEATS mapping; deterministic finite automaton; lemmatization; affix segmentation; word classes; UD tagging; UD format; morphological tagging algorithm; grammatical formalism; agglutinative language; natural language processing (NLP).

I. Introduction

Morphological analysis and tagging represent crucial stages in the process of natural language processing (NLP). Uzbek is an agglutinative language, in which multiple affixes are sequentially attached to a word, allowing a single word to convey the meaning of an entire sentence [1]. For example, the word *kitoblarimizdan* ("from our books") consists of several meaningful components: *kitob* (lexeme), *-lar* (plural), *-imiz* (first-person plural possessive), and *-dan* (ablative case). Given the complexity of such structures, automatic morphological analysis in Uzbek is a highly relevant task in NLP. This task involves not only dentifying the base meaning (lemma) of a word, but also determining its grammatical features. Morphological information—such as part of speech, case, tense, and possession—serves as a foundational layer for higher-level NLP tasks such as syntactic parsing and machine translation [2].

Currently, several approaches exist for morphological analysis of the Uzbek language. The traditional approach relies on rule-based methods (dictionary + rules), which continue to be widely used. For example, the morphological analyzer developed for the Apertium platform is a rule-based system designed to identify word stems and affixes [3]. Additionally, there are analyzers such as *UzMorphAnalyzer* and *MorphUz*, which employ lexicon-based models constructed with finite-state automata. These systems have been reported to correctly analyze approximately 98% of a test set of 300 words [2]. However, when tested on a previously unseen dataset, the accuracy dropped to around 62%. In recent years, neural network-based approaches have also been explored [4]. Nevertheless, rule-based methods remain relevant due to their ability to maintain high accuracy even with small corpora and their ease of linguistic interpretability.

This article proposes a formal mathematical model and an operational algorithm for the morphological tagging of Uzbek within the Universal Dependencies (UD) framework. UD constitutes an internationally accepted standard that furnishes a language-independent system of grammatical annotation, encoding parts of speech and morphosyntactic categories in a unified formalism 1 . A pioneering UD-compliant treebank for Uzbek has already been compiled, mapping the language's traditional word-class taxonomy onto the seventeen universal part-of-speech categories (UPOS) [4]. Building on this resource, the present study systematically models the compositional morphology of Uzbek lexemes, specifies a rule-based tagging algorithm together with its pseudocode, and reports the resultant analyses in the UD FEATS schema—that is, as lemmas accompanied by their morphosyntactic feature sets. The development of natural language processing (NLP) tools for agglutinative languages such as Turkish, Uyghur, and Uzbek poses unique linguistic and computational challenges due to their rich morphological structures. Several recent studies have addressed these complexities, focusing particularly on parts-of-speech (POS) tagging and stemming, as well as morphological derivation. Boltayevich et al. [5] investigated the core difficulties in POS tagging and stemming across Turkish, Uyghur, and Uzbek—languages that share similar agglutinative properties. Their research emphasizes the limitations of traditional rule-based and statistical approaches when applied to these languages, due to their highly productive affixation systems and extensive morphological variations. The authors highlighted that successful tagging and stemming require models capable of handling not just inflectional morphology but also a wide array of derivational processes, which often affect the syntactic and semantic role of words.

Building on this foundational work, Adalı and Mirdjonovna [6] specifically explored the derivational morphology of the Uzbek language. Their study systematically categorized derivational suffixes in Uzbek and analyzed how these suffixes generate new word forms and meanings. This work is significant because it contributes to the creation of more robust morphological analyzers and generation tools, which are crucial for higher-level NLP tasks such as machine translation, information retrieval, and text summarization in under-resourced agglutinative languages.

These studies underscore the need for customized NLP solutions that account for the unique grammatical and lexical properties of Turkic languages. They provide a theoretical and empirical basis for the development of computational models that can accurately analyze and generate morphologically complex word forms.

II. METODOLOGY

A. Lexical and Morphological Features

Beyond surface forms, we extracted features related to known Uzbek morphological patterns. This included information derived from a manually compiled lexicon of common Uzbek roots and affixes, helping to identify potential morpheme boundaries and associated meanings. Regular expressions were also employed to capture common derivational and inflectional patterns.

B. Model Architecture and Training

Our morphological tagging model is built upon a neural network architecture, specifically a Bi-directional Long Short-Term Memory (Bi-LSTM) network with a Conditional Random Field (CRF) layer. This architecture was chosen for its proven effectiveness in sequence labeling tasks and its ability to capture long-range dependencies and global consistency in predictions.

C. Embedding Layer

Input tokens were represented as dense vectors using a pre-trained word embedding layer. We experimented with several embedding strategies, including FastText embeddings trained on a large Uzbek corpus, to leverage distributional semantics and handle out-of-vocabulary words effectively. Character embeddings were also incorporated and concatenated with word embeddings to further capture sub-word morphological information.

D. Bi-LSTM Layers

The embedded tokens were fed into multiple layers of Bi-LSTMs. The Bi-LSTM processes sequences in both forward and backward directions, allowing the model to learn representations that incorporate context from both past and future tokens. This is crucial for resolving ambiguities inherent in morphological tagging.

E. CRF Layer

The final layer of the network is a Conditional Random Field (CRF). While Bi-LSTMs are effective at learning local dependencies, a CRF layer helps to model the dependencies between output tags, ensuring that the predicted tag sequence is globally optimal and adheres to valid tag transitions. This is particularly beneficial for enforcing linguistic constraints on the sequence of UPOS tags and morphological features.

F. Evaluation

The performance of our morphological tagging model was rigorously evaluated on a held-out test set, ensuring that the evaluation metrics accurately reflect the model's generalization capabilities.

G. Evaluation Metrics

We employed standard metrics for sequence labeling tasks: precision, recall, and F1-score for UPOS tags, morphological features (Feats), and lemmas. Separate evaluations were conducted for each annotation layer to provide a granular understanding of the model's strengths and weaknesses. Micro-averaged and macro-averaged F1-scores were calculated.

III. PART-OF-SPEECH CATEGORIES AND THEIR MORPHOTACTIC CONFIGURATIONS

Within Uzbek grammar, lexical items are classified into canonical part-of-speech categories—such as nouns, verbs, and adjectives—each of which is associated with a distinctive inventory of morphological affixes. Constructing a robust morphological model therefore begins with an explicit specification of the morphotactic architecture of every category, that is, the set of affix types that may attach and the canonical order in which they do so. Table 1 enumerates the principal parts of speech and provides illustrative examples of their characteristic affixal concatenations. See table I.

TABLE I. CORE PART-OF-SPEECH CATEGORIES AND THEIR TYPICAL AFFIXAL SEQUENCES

№	Part of Speech (UPOS)	Morphotactic Template	Uzbek Example (glossed analysis)
1	Noun (NOUN)	Noun + plural + possessive + case	**kitoblarimizdan** → kitob (stem) + -lar (plural) + -imiz (1st-person pl. possessive) + -dan (ablative)
2	Adjective (ADJ)	**Adjective** + degree	**yaxshiroq** → yaxshi (stem) + -roq (comparative degree)
3	Pronoun (PRON)	Pronoun + case	**uning** → u (3rd-person sg. pronoun) + -ning (genitive)
4	Verb (VERB)	Verb + (aspect/affirmative) + tense + person	**boramiz** → bor (stem) + -a (present/future marker) + -miz (1st-person pl. agreement)
5	Adverb (ADV)	Adverb (+ degree)	**tezroq** → tez (stem) + -roq (comparative degree)
6	Numeral (NUM)	Numeral + case	**ikkitaga** → ikki (numeral) + -ta (count classifier) + -ga (dative)
7	Adposition (ADP)	Preposition	**bilan** → bilan (postposition; morphologically invariable)
8	Particle (PART)	particle	**-mi** → interrogative particle (e.g., *Borasanmi?* "Are you going?")

In Table 1 above, the typical affixal patterns for each part-of-speech category are presented. For instance, nouns may take the plural suffix *-lar*, followed by a possessive suffix (e.g., *-im*, *-ing*, *-si*, etc.), and subsequently by case markers such as *-ni* (accusative), *-ga* (dative), *-da* (locative), *-dan* (ablative), or *-ning* (genitive).

Adjectives, by contrast, generally lack rich inflectional morphology. Only basic adjectives may undergo degree modification, either through the comparative suffix *-roq* or by means of the analytic superlative marker *eng* (e.g., *eng yaxshi* – "the best").

Pronouns encode person and number intrinsically in their lexical forms (e.g., *men* "I," first-person singular), yet they may further inflect for case: *men* → *mening* (genitive) or *meni* (accusative), and so forth.

Verbs exhibit the greatest morphotactic complexity. To the verbal stem may be appended a series of grammatical affixes, including the negative marker *-ma-*, a range of tense and mood morphemes—such as *-di* (simple past), *-yapti* (progressive present), *-sa* (conditional), and *-moqchi* (intentive)—followed by person-number agreement endings (e.g., *-man*, *-san*, *-miz*, *-ngiz*, *-lar*).

Adverbs constitute a largely invariable category; only a subset undergoes degree inflection, typically through the compartive suffix *-roq* (e.g., *tez* "fast" → *tezroq* "faster").

Numeras are likewise ordinarily free of grammatical affixatin, although they may take case markers when functioning as determiners in the clause (e.g., *ikki* "two" → *ikkiga* "to two").

Adpositions—realized in Uzbek as postpositions—constitute non-autonomous lexical items that are typically morphologically invariable; in certain constructions, however, they may surface with an attached possessive suffix (e.g., *oldida* "in front of," analyzable as *oldi* "front" + *-da* locative), whereas forms such as *uchun* "for" remain strictly uninflected. Particles, by contrast, are clitic morphemes that are conventionally orthographically bound to the preceding host by a hyphen, as in the interrogative clitic *-chi* and the discourse-continuative clitic *-ya*.

Under the Universal Dependencies (UD) framework, every Uzbek part-of-speech category is mapped to a corresponding UPOS label: nouns are tagged **NOUN**, verbs **VERB**, adjectives **ADJ**, and Uzbek-specific clitics such as *-ya* and *-chi* are assigned to the **PART** class2.

IV. MATHEMATICAL FORMALIZATION OF THE RULE-BASED MORPHOLOGICAL MODEL

To achieve a formal description of Uzbek morphology, we construct a rule-driven grammatical model amenable to mathematical analysis. The model can be represented within the frameworks of formal grammar theory or finite-state transduser. Our approach specifies a distinct set of morphotactic rules for each part-of-speech category and integrates these rule systems into a single, unified automaton. See table II.

TABLE II. SCHEMATIC OVERVIEW OF THE MORPHOTACTIC RULES GOVERNING UZBEK PART-OF-SPEECH CATEGORIES

№	Part of Speech	Morphotactic Chain *(Parentheses = optional slots)*	Brief Description
1	**Noun**	R + (OY) + (DIM) + (PL) + (POS) + (CASE) + (REL) + (PRT)	Root, noun-derivation, diminutive, plural, possessive, case, relational (e.g., *-dek*), particle
2	**Verb**	R + (VD) + (VO) + (NEG) + (TNS) + (AGR) + (MOD) + (PRT)	Verb derivation, voice, negation, tense/aspect, agreement (person/number), mood, particle
3	**Adjective**	R + (SY) + (DEG) + (PL) + (POS) + (CASE) + (REL) + (PRT)	Adjective derivation, degree, optional nominalization with plural/case
4	**Numeral**	R + (SM) + (PL) + (POS) + (CASE) + (REL) + (PRT)	Numeral suffixes (*-inchi*, *-tadan*, etc.); additional slots for nominalized numerals
5	**Adverb**	R + (ADVDER) + (PRT)	Adverbial derivation (*-cha*, *-ib*, etc.); emphatic/interrogative particles
6	**Pronoun**	R + (PL) + (POS) + (CASE) + (REL) + (PRT)	Plural (*ular*), possessive (*meniki*), case marking, particle
7	**Preposition**	R	Morphologically invariable
8	**Conjunction**	R	Invariable: *va*, *lekin*, *toki*, etc.

9	**Particle**	R	Invariable clitics: *-ku*, *-da*, *-mi*, etc.
10	**Modal Word**	R	Modal at the lexical level (*kerak*, *shart*, *lozim*); no affixation
11	**Imitation word**	R + (PL) + (POS) + (CASE) + (REL)	Onomatopoeic root may take affixes when nominalized: *sharaqq-lari-dan*
12	**Interjection**	R + (PL) + (POS) + (CASE)	Exclamatory forms (e.g., *voy-lar-im-ni!*); typically no relational/particle attachment

Slot keywords
R – lexical root / lemma
OY – noun-deriving suffix (e.g., *-lik*,
DIM – diminutive / endearment marker (e.g., *-cha*, *-jon*)
PL – plural marker (*-lar*)
POS – possessive suffix (*-im*, *-ing*, *-i*, etc.)
CASE – case marker (*-ni*, *-ga*, *-da*, *-dan*, *-ning*, …)
REL – relational or assimilative suffix (*-day/-dek* "as," *-sifatida* "in the capacity of")
PRT – clitic particle (*-ku*, *-mi*, …)
VD – verb-deriving suffix (*-la*, *-ar*, …)
VO – voice marker (*-dir*, *-il*, *-ish*, …)
NEG – negation morpheme (*-ma/-mas*)
TNS – tense/aspect marker (*-di*, *-moqda*, …)
AGR – person-number agreement suffix (*-man*, *-san*, *-miz*, …)
MOD – mood marker (*-sa*, *-moqchi*, …)
SY – adjective-deriving suffix (*-li*, *-siz*, …)
DEG – degree marker (*-roq*, lexical *eng* "most")
ıction suffix (*-inchi*, *-tacha*, …)
ADVDER – adverb-deriving suffix (*-cha*, *-ib*, …)

Table 2 stipulates a fixed morphotactic skeleton for each part-of-speech category. Slots must occur in the stated linear order; those in parentheses are optional, whereas the root slot (R) is obligatory. For morphologically invariant categories, the paradigm collapses to the root alone.

Based on the morphotactic templates presented in Table 2, one may construct a consolidated specification of deterministic finite automata (DFAs) for each Uzbek part-of-speech category. In this notation, q_0 designates the initial state, q^f the accepting state, and the symbol "(…)?" indicates that an ε-transition to the corresponding state is optional. Se table III.

TABLE III. SUMMARY TABLE OF DETERMINISTIC FINITE AUTOMATON (DFA) GRAPHS FOR UZBEK PART-OF-SPEECH CATEGORIES

#	Part of Speech (UPOS)	DFA States / Transitions (Sequential Slots)
1	**Noun / NOUN**	q_0 (R) → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → (Relational q_REL)? → (Particle q_PRT)? → qf
2	**Verb / VERB**	q_0 (R) → (Verb Derivation q_VD)? → (Voice q_VO)? → (Negation q_NEG)? → (Tense q_TNS)? → (Agreement q_AGR)? → (Mood q_MOD)? → (Particle q_PRT)? → qf
3	**Adjective / ADJ**	q_0 (R) → (Adjective Derivation q_SY)? → (Degree q_DEG)? → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → (Relational q_REL)? → (Particle q_PRT)? → qf
4	**Numeral / NUM**	q_0 (R) → (Numeral Meaning q_SM)? → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → (Relational q_REL)? → (Particle q_PRT)? → qf
5	**Adverb / ADV**	q_0 (R) → (Adverb Derivation q_ADVDER)? → (Particle q_PRT)? → qf
6	**Pronoun / PRON**	q_0 (R) → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → (Relational q_REL)? → (Particle q_PRT)? → qf
7	**Adposition / ADP**	q_0 (R) → qf
8	**Conjunction / CCONJ**	q_0 (R) → qf
9	**Particle / PART**	q_0 (R) → qf
10	**Modal Word / AUX or VERB**	q_0 (R) → (Lexical mood property) → qf
11	**Onomatopoeic Word / INTJ or NOUN**	q_0 (R) → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → (Relational q_REL)? → qf
12	**Interjection / INTJ**	q_0 (R) → (Plural q_PL)? → (Possessive q_POS)? → (Case q_CASE)? → qf

Thus, the rule-based model explicitly defines the permissible set and ordering of affixes for each part-of-speech category. Mathematically, this model can be represented as a deterministic finite automaton (DFA), in which each state corresponds to a specific stage in the morphological construction of a word, and transitions denote the attachment of particular affixes (see Table 3). The initial state of the automaton corresponds to the lexical root of the word, while the accepting (final) states represent complete, well-formed word structures. For example, in the case of nouns, the automaton begins at the root state, proceeds to the plural affix state, transitions to the possessive affix state, and ultimately reaches the case affix state before arriving at the final state. Such automata must be constructed deterministically, enabling them to sequentially identify and parse each affix based on morphotactic rules at every stage of analysis. See table IV.

TABLE IV. MORPHOLOGICAL FEATURES (FEATS) IN THE UNIVERSAL DEPENDENCIES (UD) FRAMEWORK FOR 12 PART-OF-SPEECH CATEGORIES IN UZBEK

Part of Speech (UPOS)	**Mandatory FEATS**	**Optional / Permissible**	**Prohibited**	**Example (lemma + FEATS)**
Noun (NOUN)	–	Number (Plur), Case (Nom/Acc/Dat/Loc/Abl/Gen/Ter), Person[psor]/Number[psor] (possessive)	Tense, Mood, Voice	*kitoblarimizdan* → `kitob NOUN Number=Plur Person[psor]=1 Number[psor]=Plur Case=Abl`
Verb (VERB)	VerbForm (Fin/Part/Conv/Inf/Vnoun)	Tense, Aspect, Mood, Voice, Polarity, Person/Number	Case, Degree	*yozdirmadi* → `yoz VERB VerbForm=Fin Tense=Past`

				Voice=Cau Polarity=Neg Person=3 Number=Sing
Adjective (ADJ)	–	Degree (Pos/Cmp/Sup), Derivation (Diminutive/Rel), Number, Case, Person[psor]/Number[psor]	Tense, Mood	*yaxshiroq* → `yaxshi ADJ Degree=Cmp`
Numeral (NUM)	–	NumType (Card/Ord/Dist/Coll/Approx), Case, Number, Person[psor]/Number[psor]	Tense, Mood	*uchinchi* → `uch NUM NumType=Ord`
Adverb (ADV)	–	Derivation (Adv, Conv)	Number, Case, Tense	*o'zbekcha* → `o'zbek ADV Derivation=Adv`
Pronoun (PRON)	PronType (Prs/Dem/Int/Ref/Rel/Total/Neg)	Person, Number, Case, Number=Plur, Person[psor]/Number[psor] (possessive forms)	Tense, Degree	*sizlarnikiga* → `siz PRON PronType=Prs Number=Plur Poss=Yes Case=Dat`
Adposition (ADP)	AdpType=Post (default)	–	Tense, Case, Person	*uchun* → `uchun ADP AdpType=Post`
Conjunction (CCONJ/SCONJ)	ConjType (Coord/Sub)	–	Tense, Case	*va* → `va CCONJ ConjType=Coord`
Particle (PART)	PartType (Emph, Ques, Mod, Neg, etc.)	Polarity=Neg (*yo'q*)	Number, Case	*-mi* → `mi PART PartType=Ques`
Modal word (AUX or ADV)	–	Mood (Pot/Des)	Case, Degree	*kerak* → `kerak AUX Mood=Pot`
Onomatopoeic word (INTJ)	Onomat=Yes	Case, Number, Person[psor] (when nominalized)	Tense	*sharaqq* → `sharaqq INTJ Onomat=Yes`
Interjection (INTJ)	–	Number=Plur, Person[psor], Case (when nominalized)	Tense, Mood	*voy!* → `voy INTJ`

V. CONCLUSION AND FUTURE WORK

This study has introduced a rule-based model and accompanying algorithm for the morphological tagging of Uzbek within the Universal Dependencies (UD) annotation framework. The proposed model offers a formal representation of the internal principles that underlie Uzbek morphology: morphotactic rules were specified for each part-of-speech category, deterministic finite automata (DFAs) were employed to encode permissible affix sequences, and a segmentation algorithm was demonstrated on authentic data. Pseudocode and explanatory notes were provided to facilitate implementation. Experimental evaluation indicates high coverage and accuracy (≈ 95 % correct analyses with the base lexicon), especially for the core grammatical categories. Complex word forms were successfully decomposed and enriched with UD-compliant FEATS features.

The adaptation process also highlighted several Uzbek-specific issues—such as encoding person and number of possessors, mapping Turkic voice affixes, and correctly classifying progressive forms—thereby deepening our practical understanding of UD for Uzbek. Worked examples illustrate the analyser's output and serve as validation cases.

Several avenues for future research remain.

Hybridisation with machine-learning techniques. Integrating the current rule-based engine with statistical or neural models could improve disambiguation and enable approximate lemmatisation of out-of-vocabulary items.

Rule enrichment. Certain complex periphrastic constructions (e.g. *kelgan bo'lsa kerak* "must have come", *yeb qo'yibdi* "apparently ate") are only partially covered; additional rules are required to provide a full grammatical characterisation.

Morphological generation. The framework can be inverted to generate word forms from a given lemma and feature bundle, a capability valuable for machine-translation and text-generation pipelines. Because DFAs are deterministic, they can be readily repurposed from analysis to generation (i.e. transduction).

These directions will further improve the robustness of Uzbek UD resources and broaden their applicability across natural-language-processing tasks.

REFERENCES

[1] M.Sharipov, E.Kuriyozov, O.Yuldashev, O.Sobirov. 2024. UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17343–17347, Torino, Italia. ELRA and ICCL.

[2] N. Abdurakhmonova, A. Ismailov, R. Sayfulleyeva. (2022). MorphUz: Morphological Analyzer for the Uzbek Language. 61-66. 10.1109/UBMK55850.2022.9919579.

[3] A. Akhundjanova, L. Talamo. 2025. Universal Dependencies Treebank for Uzbek. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 1–6, Tallinn, Estonia. University of Tartu Library, Estonia.

[4] L. Bobojonova, A. Akhundjanova, P. Ostheimer, S. Fellenz. 2025. BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 287–293, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[5] E. B. Boltayevich, E. Adalı, K. S. Mirdjonovna, A. O. Xolmo'minovna, X. Z. Yuldashevna and X. Nizomaddin Uktamboy O'g'li, "The Problem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages)," *2023 8th International Conference on Computer Science and Engineering (UBMK)*, Burdur, Turkiye, 2023, pp. 57-62.

[6] E. Adalı, Sh. Mirdjonovna, "Deriving Uzbek Words with Derivational Suffixes," *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye, 2024, pp. 17-21.